จุฬารัตน์ ตันประเสริฐ

ซอฟต์แวร์ไทยโอซีอาร์ : พนักงานพิมพ์ดีดคนใหม่

รายละเอียดของซอฟต์แวร์อ่านไทย และเจาะลึกเรื่องราวของเทคโนโลยีโอซีอาร์ให้ทราบกันก่อน

สำหรับท่านที่ใช้เวลาส่วนใหญ่อยู่กับการพิมพ์งานและจัดเตรียมแก้ไขเอกสารด้วยเครื่องพิมพ์ดีด หรือด้วยซอฟต์แวร์ประมวลคำ (Word Processing Software) ก็ตาม เคยนึกบ้างไหมว่าถ้าเราไม่ต้องพิมพ์เองก็ดีน่ะสิ ทางหนึ่งคือถ่ายเทงานให้คนอื่นทำแทน ข้อนี้ก็แล้วแต่เทคนิคของแต่ละท่าน ว่าจะเอาอะไรมาหลอกล่อให้ผู้อื่นทำแทนได้สำเร็จ แต่ในปัจจุบันนี้ไม่ต้องเอาเปรียบผู้อื่นอีกต่อไป มีเทคโนโลยีใหม่ที่เรียกว่าโอซีอาร์เกิดขึ้นแล้ว สำหรับภาษาไทย (สำหรับภาษาอังกฤษนั้นเกิดมานานหลายสิบปีแล้ว) ซอฟต์แวร์ประเภทนี้ของภาษาไทยเรานั้นหรือที่มักจะเรียกกันง่าย ๆ ว่าซอฟต์แวร์ "ไทยโอซีอาร์" เกิดขึ้นมา 3 ซอฟต์แวร์แล้วในเวลาใกล้เคียงกันตั้งแต่กลางปี พ.ศ. 2539 เป็นต้นมา ซอฟต์แวร์ทั้ง 3 นี้คือ

1. ซอฟต์แวร์ไทยโอซีอาร์ (ThaiOCR) พัฒนาโดย บริษัท เอเทรียม เทคโนโลยี จำกัด

2. ซอฟต์แวร์อ่านไทย พัฒนา Thai & English OCR Engine โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ และพัฒนาส่วนติดต่อกับผู้ใช้ (Interface) โดย บริษัท ไทยซอฟท์ จำกัด

3. ซอฟต์แวร์ทีเร็ค (T-Rec) ซึ่งพัฒนาโดย บริษัท เอ็นเอสที อิเลคทรอนิค พับลิชชิ่ง จำกัด

ซอฟต์แวร์ทั้งสามนี้ทำงานเฉพาะกับตัวพิมพ์ภาษาไทยและภาษาอังกฤษเท่านั้น ในปัจจุบันนี้ยังไม่มีซอฟต์แวร์ไทยโอซีอาร์ในท้องตลาดที่ทำงานได้กับตัวเขียนไทยเลย สำหรับซอฟต์แวร์ทั้งสามข้างต้นนี้ ท่านสามารถหาซื้อมาลองใช้กันได้แล้วตามร้านจำหน่ายสินค้าไอทีทั่วไป (ระวังสินค้าปลอมด้วยนะจะบอกให้) ความสามารถของแต่ละซอฟต์แวร์ก็จะแตกต่างกันไป อัตราการรู้จำตัวอักษรที่ถูกต้องนั้นจะขึ้นอยู่กับฟอนต์ที่ใช้ในเอกสารที่นำมาใช้งาน และความสะอาดของเอกสารต้นฉบับด้วย และเนื่องจากซอฟต์แวร์ประเภทนี้รับอินพุตในรูปของภาพลักษณ์ของเอกสาร (Document Image) ซึ่งรับมาโดยผ่านเครื่องสแกนเนอร์หรือเครื่องกราดตรวจ (ตามที่พจนานุกรมศัพท์คอมพิวเตอร์ ฉบับราชบัณฑิตยสถานกำหนดไว้) นั่นคือท่านผู้ใช้ควรจะต้องมีเครื่องกราดตรวจด้วยในการที่จะใช้งานซอฟต์แวร์นี้ ถ้าไม่มีท่านก็อาจจะไปหยิบยืมใช้ของคนอื่นและนำแฟ้มภาพลักษณ์เหล่านั้นมาใช้กับซอฟต์แวร์อีกทีก็ได้

แต่เนื่องจากการพึ่งพาอาศัยกันระหว่างซอฟต์แวร์นี้กับเครื่องกราดตรวจ ส่งผลให้มีการนำซอฟต์แวร์ประเภทนี้มาแถมให้กับผู้ซื้อเครื่องกราดตรวจ ซึ่งเมื่อท่านเลือกซื้อเครื่องกราดตรวจและอยากจะได้ซอฟต์แวร์ประเภทนี้ติดมาด้วย ก็ถามรายละเอียดจากผู้ขายให้ดีก่อนว่ามีแถมหรือเปล่า ถ้าแถม มีแถมให้กับเครื่องรุ่นใดบ้าง และซอฟต์แวร์ที่แถมมานั้นชื่ออะไร และเวอร์ชันอะไร จะได้ประหยัด (ตามนโยบายของประเทศเราในยุค IMF) ไม่ต้องควักกระเป๋าจ่ายให้กับซอฟต์แวร์อีกรอบ

เรื่องน่าอ่าน

ซอฟต์แวร์ไทยโอซีอาร์
: พนักงานพิมพ์ดีดคนใหม่

สำหรับบทความนี้จะกล่าวถึงรายละเอียดของซอฟต์แวร์อ่านไทยให้ได้ทราบกัน และจะไม่กล่าวถึงรายละเอียดของซอฟต์แวร์ประเภทเดียวกันตัวอื่น ทั้งนี้ก็เพราะว่าไม่ทราบรายละเอียดของการพัฒนาของซอฟต์แวร์เหล่านั้น ก่อนที่จะเจาะลึกเรื่องราวของซอฟต์แวร์อ่านไทย ผู้เขียนขอเล่าย้อนถึงอดีตของเทคโนโลยีโอซีอาร์ให้ผู้อ่านทราบกันก่อน จะได้เห็นภาพและเข้าใจว่ากว่าจะได้มาซึ่งเทคโนโลยีไทยโอซีอาร์เนี่ย! ยากนะ

ก่อนอื่นขออธิบายคำว่าโอซีอาร์ให้กระจ่างแจ้งก่อน คำเต็ม ๆ นั้นคือ Optical Character Recognition (OCR) ซึ่งก็แปลเป็นไทยได้ว่าการรู้จำอักขระด้วยแสง (จากพจนานุกรมคอมพิวเตอร์สำหรับเยาวชน โดย ดร. ครรชิต มาลัยวงศ์) อินพุตของเทคนิคนี้คือภาพลักษณ์ของเอกสารที่ได้จากการสแกนหรือการกราดตรวจ และเอาต์พุตคือแฟ้มข้อมูล (Text File) ซึ่งทำงานเสมือนเป็นพนักงานพิมพ์ดีดนั่นเอง เทคนิคนี้เมื่อได้รับอินพุตจะทำตัดแยกแต่ละบรรทัด และแต่ละภาพลักษณ์ของตัวอักษรออกจากกัน หลังจากนั้นก็จะพิจารณาว่าภาพลักษณ์แต่ละตัวคือตัวอักษรใด เมื่อได้ครบทุกภาพลักษณ์ตัวอักษรแล้วก็จะนำมาประกอบกันใหม่รวมเป็นแฟ้มข้อความ ซึ่งระบบคอมพิวเตอร์สามารถเข้าใจได้แล้วว่าบรรทัดใด ย่อหน้าใดมีอักษรใดอยู่บ้าง ผู้ใช้สามารถแก้ไข เพิ่มเติมข้อความได้โดยใช้โปรแกรมประมวลคำใด ๆ ก็ได้

ข้อเด่นของเทคโนโลยีโอซีอาร์คือ

1. ประหยัดทรัพยากรมนุษย์ที่ต้องใช้ในการพิมพ์ และประหยัดเวลา โดยทั่วไปแล้วซอฟต์แวร์ประเภทโอซีอาร์จะทำงานได้เร็วกว่าการพิมพ์ของมนุษย์หลายเท่าตัวทีเดียว (จะกี่เท่าก็ขึ้นอยู่กับว่าเป็นซอฟต์แวร์ใด เครื่องรุ่นใด และความเร็วของพนักงานพิมพ์ดีดที่นำมาเปรียบเทียบ)

2. ประหยัดเนื้อที่ในการจัดเก็บข้อมูล เนื่องมาจากการเก็บข้อมูลในรูปกระดาษนั้นใช้เนื้อที่มาก เก็บกันทีเป็นห้อง ๆ ทีเดียว เราก็พยายามแก้ไขกันโดยนำเอกสารเหล่านั้นมาสแกนเก็บเป็นแฟ้มภาพลักษณ์ของหน้าเอกสาร ซึ่งใช้เนื้อที่น้อยลงเมื่อเทียบกับการเก็บในรูปแบบกระดาษ และยังปลอดภัยจากพวกปลวก หนู และแมลงสาบมากขึ้นอีกด้วย

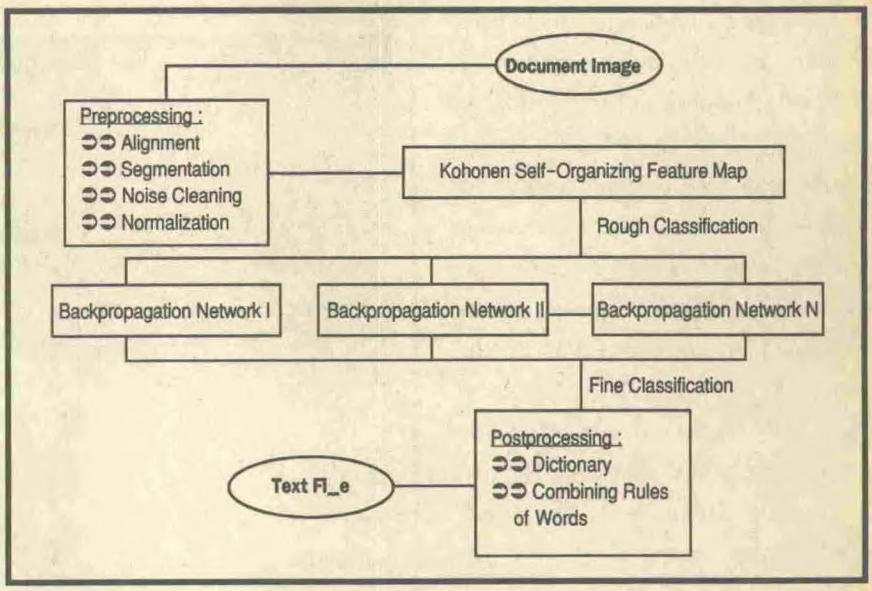
อย่างไรก็ตาม การพัฒนาก็ยังคงดำเนินต่อไป มีการพัฒนาโอซีอาร์ขึ้นมาแปลงภาพลักษณ์ของเอกสารเป็นแฟ้มข้อมูลอีกหนึ่งขั้นตอน ซึ่งขนาดของแฟ้มข้อมูลที่ได้นั้นจะเล็กลงไปเป็นอย่างมาก ตัวอย่างเช่น ผู้เขียนลองสแกนเอกสารขนาด A4 ด้วยความละเอียดระดับ 300 dpi แบบขาวดำ แล้วจัดเก็บในรูปแบบ ".BMP" ปรากฏว่าแฟ้มภาพลักษณ์มีขนาด 672 KB และเมื่อแปลงเป็นแฟ้มข้อความแล้วจัดเก็บในรูปแบบ ".TXT" แฟ้มข้อความมีขนาดเหลือเพียง 3KB เท่านั้น ซึ่งเห็นได้ว่าประหยัดเนื้อที่ในการจัดเก็บได้เป็นร้อยเท่าทีเดียว และยังส่งผลให้ผู้ใช้สามารถค้นหาข้อมูลได้อย่างสะดวกและรวดเร็วขึ้นโดยการใช้เทคนิค Full Text Search ได้อีกด้วย ซึ่งจะเห็นได้ว่าโอซีอาร์เป็นเทคโนโลยีที่มีความสำคัญมากกับการจัดการงานเอกสาร

จากหลักฐานทางประวัติศาสตร์ เท่าที่ค้นพบกัน สิทธิบัตรของโอซีอาร์ของโลกตามหลักฐานที่ค้นพบ ชิ้นแรกพัฒนาในประเทศเยอรมันโดย จี ทัวเชค (G. Tausheck) ในปี พ.ศ. 2472 จดสิทธิบัตรในประเทศสหรัฐอเมริกา เลขที่ 2 026 329 ในปี พ.ศ. 2478 เรียกว่าเครื่องอ่าน (Reading Machine) และในเวลาใกล้เคียงกันสิทธิบัตรที่สองค้นพบในประเทศสหรัฐอเมริกา โดย พี ดับบิว แฮนเดล (P. W. Handel) ในปี พ.ศ. 2476 และจดสิทธิบัตรในปีนั้นเลย โดยเป็นสิทธิบัตรในประเทศสหรัฐอเมริกา เลขที่ 1 915 993 เรียกว่าเครื่องจัดทำสถิติ (Statistical Machine) ในระยะเริ่มแรกนั้นนักวิจัยฝันกันว่าจะสามารถผลิตเครื่องอ่านตัวหนังสือที่ดีมีประสิทธิภาพได้โดยง่าย แต่เมื่อลงมือค้นคว้ากันแล้ว จึงรู้ว่ากว่าฝันของพวกเขาจะเป็นจริงขึ้นมาได้ ต้องใช้เวลาและความพยายามกันอีกมากทีเดียว

ในสมัยนั้นยังไม่มีการพัฒนาเครื่องคอมพิวเตอร์ขึ้นมาใช้งาน วิธีที่ทัวเชคใช้ในการรู้จำตัวอักษรคือการจับคู่แผ่นแบบ/ตัวพราง (Template/Mask Matching) พื้นฐานของเทคนิคคือภาพลักษณ์ตัวอักษรจะถูกกราดตรวจ (scan) ด้วยตัวตรวจจับลำแสง (Photodetector) ผ่านตัวพราง (Mask) ซึ่งมีลำแสงยิงผ่าน และเมื่อผ่านตัวพรางที่จับคู่ได้ดีกับภาพลักษณ์ตัวอักษร ตัวตรวจจับลำแสงจะรู้จำภาพลักษณ์ตัวอักษรได้แล้วส่งตัวอักษรนั้นออกพิมพ์บนกระดาษ วิธีนี้ได้รับการพัฒนาให้สอดคล้องกับเทคโนโลยีใหม่ ๆ ในเวลาต่อมาในชื่อ "วิธีการจับคู่แผ่นแบบ" (Template Matching Methods) อย่างไรก็ตาม เทคนิคในกลุ่มนี้มีประสิทธิภาพในการรู้จำได้ดีเฉพาะกับตัวอักษรแบบตัวพิมพ์เท่านั้น ส่วนกลุ่มตัวอักษรแบบเขียนซึ่งมีความแปรปรวนในขนาดของรูปร่างของตัวอักษรเป็นอย่างมาก เนื่องจากมนุษย์มักจะเขียนตัวอักษรแตกต่างกัน ดังนั้นจึงมีกลุ่มนักวิจัยคิดค้นเทคนิคใหม่ที่เหมาะสมกับการรู้จำตัวอักษรแบบเขียน (Handwritten Character Recognition) เรียกว่า "วิธีการวิเคราะห์โครงสร้าง" (Structure Analysis Method) ขึ้น หลักการคือการแบ่งโครงสร้างออกเป็นส่วนย่อย แล้วศึกษาว่าส่วนย่อยแต่ละส่วนนั้นประกอบด้วยลักษณะเด่น (Features) อย่างไรบ้าง พร้อมทั้งศึกษาความสัมพันธ์ระหว่างส่วนย่อยด้วย ซึ่งลักษณะของแต่ละส่วนย่อยและความสัมพันธ์ระหว่างส่วนย่อยจะถูกนำมาใช้ในการระบุตัวอักษร ซึ่งจะเห็นได้ว่าการคัดลักษณะเด่น (Feature Extraction) เป็นหัวข้อที่สำคัญมากสำหรับวิธีการวิเคราะห์โครงสร้าง

สองวิธีการดังกล่าวเป็นวิธีการหลักที่ใช้ในการรู้จำตัวอักษร มีข้อเด่นข้อด้อยของแต่ละวิธี อาทิเช่น วิธีการจับคู่แผ่นแบบนั้นจะไวต่อการเปลี่ยนตำแหน่งเป็นอย่างมากแต่การจับคู่ในภาพรวมดีมาก ซึ่งตรงกันข้ามกับวิธีการ

วิเคราะห์โครงสร้างซึ่งสามารถจับลักษณะเด่นของเส้นย่อยที่ประกอบเป็นตัวอักษรได้ ดังนั้นจึงมีงานวิจัยบางส่วนพยายามนำสองวิธีการนี้มาพัฒนาใช้ร่วมกันเพื่อหวังที่จะดึงข้อเด่นของทั้งสองวิธีการออกมาใช้งานเพื่อให้เกิดประโยชน์สูงสุด ซึ่งทั้งสองวิธีนี้เป็นเทคนิคที่ใช้สำหรับการพัฒนาเทคโนโลยีโอซีอาร์มาตั้งแต่ระยะเริ่มแรก และเรื่อยมาจนถึงปัจจุบัน อย่างไรก็ตามจากการเจริญเติบโตอย่างรวดเร็วของเทคโนโลยีในทุก ๆ ด้าน ส่งผลให้มีความพยายามที่จะนำเทคนิคใหม่ ๆ มาใช้ในการทำโอซีอาร์ และเทคนิคที่กำลังได้รับความนิยมเป็นอย่างสูงคือโครงข่ายประสาทเทียม (Artificial Neural Networks) ซึ่งเป็นความพยายามในการหาเทคนิคที่ทำให้ระบบคอมพิวเตอร์สามารถทำงานบางอย่างได้เช่นเดียวกับมนุษย์



รูปที่ 1

กว่าต่างประเทศจะพัฒนาเทคโนโลยีโอซีอาร์สำหรับภาษาอังกฤษมาได้ ณ สถานะปัจจุบันเขาใช้เวลากันนานกว่า 70 ปี และในปัจจุบันซอฟต์แวร์ประเภทนี้สำหรับภาษาอังกฤษมีอยู่มากมายหลายเจ้าทีเดียว อาทิเช่น Omnipage, TypeReader, Prime Recognition, และ Readiris เป็นต้น ทุก ๆ ซอฟต์แวร์จะบอกว่าอัตราการรู้จำตัวอักษรอยู่ที่ 99.99 เปอร์เซ็นต์เกือบทุกยี่ห้อ ถ้าท่านผู้อ่านสนใจลองค้นหาข้อมูลเพิ่มเติมได้ในระบบอินเทอร์เน็ต สำหรับเทคโนโลยีนี้ของภาษาไทย ได้มีคณาจารย์และนักวิจัยในมหาวิทยาลัยให้ความสนใจค้นคว้าและพัฒนามากว่า 10 ปีแล้ว ตัวอย่างงานวิจัยเช่น สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง นำทีมโดย ดร. ชม กิ้มปาน สถาบันบัณฑิตพัฒนบริหารศาสตร์ นำโดย ดร. พิพัฒน์ หิรัญวนิชกร และจุฬาลงกรณ์มหาวิทยาลัย โดย ดร. ชิดชนก เหลือสินทรัพย์ และ ดร. สมชาย จิตะพันธ์กุล เป็นต้น

การพัฒนาซอฟต์แวร์ไทยโอซีอาร์ดูเหมือนว่าจะยากกว่าการพัฒนาโอซีอาร์สำหรับภาษาอังกฤษ ด้วยลักษณะพิเศษหลาย ๆ อย่างของตัวอักษรไทย อาทิเช่น มีตัวอักษรที่มีหน้าตาคล้ายกันหลายกลุ่ม เช่น "ก ภ ถ" "บ ป พ ฟ ผ ฝ" "ข ฃ ช ฆ" ประโยคในภาษาไทยประกอบไปด้วย 4 ระดับได้แก่ ระดับวรรณยุกต์ ระดับสระบน ระดับตัวอักษร และระดับสระล่าง นอกจากนี้จำนวนตัวอักษรทั้งหมดของภาษาไทยยังมีจำนวนมากกว่าตัวอักษรของภาษาอังกฤษอีกด้วย การศึกษาและวิจัยงานไทยโอซีอาร์นั้นอาจจะแบ่งเป็นการศึกษาไทยโอซีอาร์สำหรับตัวพิมพ์หรือสำหรับตัวเขียน สำหรับตัวพิมพ์ก็อาจจะศึกษาเฉพาะฟอนต์เดียว หรือศึกษาเทคนิคที่ใช้ได้กับหลาย ๆ ฟอนต์พร้อมกัน นอกจากนี้ยังแบ่งได้อีกว่าเป็นการรู้จำแบบในสาย (on-line) หรือนอกสาย (off-line) ซึ่งเทคนิคที่พัฒนาก็จะแตกต่างกันไป เรื่องที่ต้องศึกษาประกอบด้วย

1. การกำจัดสิ่งสกปรกออกจากภาพลักษณ์ของหน้าเอกสาร (Noise Cleaning)

2. การตัดแยกบรรทัดและภาพลักษณ์ของแต่ละตัวอักษร (Segmentation)

3. การปรับแต่งภาพลักษณ์ (Normalization)

4. การคัดเลือกลักษณะเด่น (Feature Extraction)

5. การทำให้ภาพลักษณ์ตัวอักษรบาง (Thining)

6. การทำให้ภาพลักษณ์ตัวอักษรมัว (Blurring)

7. เทคนิคการรู้จำ เช่น Template Matching, Structure Analysis หรือ Neural Networks เป็นต้น

8. วิธีการตรวจสอบคำผิดของผลลัพธ์ที่ได้จากการรู้จำ (ไม่จำเป็นว่าซอฟต์แวร์โอซีอาร์ที่พัฒนาขึ้นมาแต่ละซอฟต์แวร์จะต้องใช้ทุก ๆ วิธีการข้างต้น)

ถ้าท่านผู้อ่านท่านใดสนใจในรายละเอียดของแต่ละวิธี ลองหาอ่านกันดูได้ในรายการเอกสารอ้างอิงที่แสดงในตอนท้ายของบทความนี้ ถึงตอนนี้ก็จะขอกล่าวถึงรายละเอียดในการพัฒนาซอฟต์แวร์อ่านไทย ซึ่งเวอร์ชัน 1.0 ออกสู่ท้องตลาดตั้งแต่ต้นปี พ.ศ. 2540 และเวอร์ชัน 1.5 กำลังจะออกสู่สายตาทุก ๆ ท่านในเร็ว ๆ วันนี้

การทำงานของซอฟต์แวร์แบ่งได้เป็น 3 ขั้นตอนด้วยกันคือ

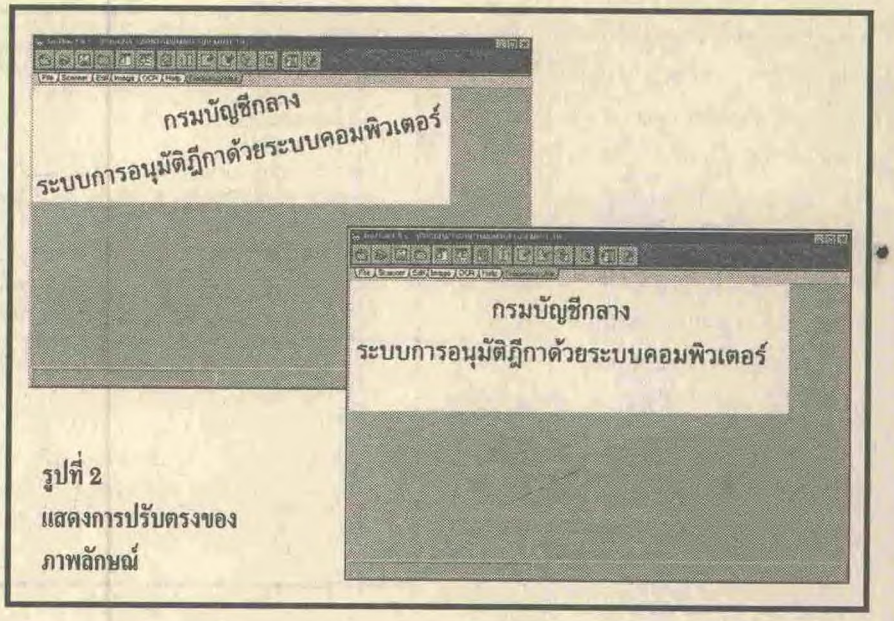
1. ขบวนการเตรียมประมวลผล (Preprocessing)

2. ขบวนการรู้จำ (Recognition Process)

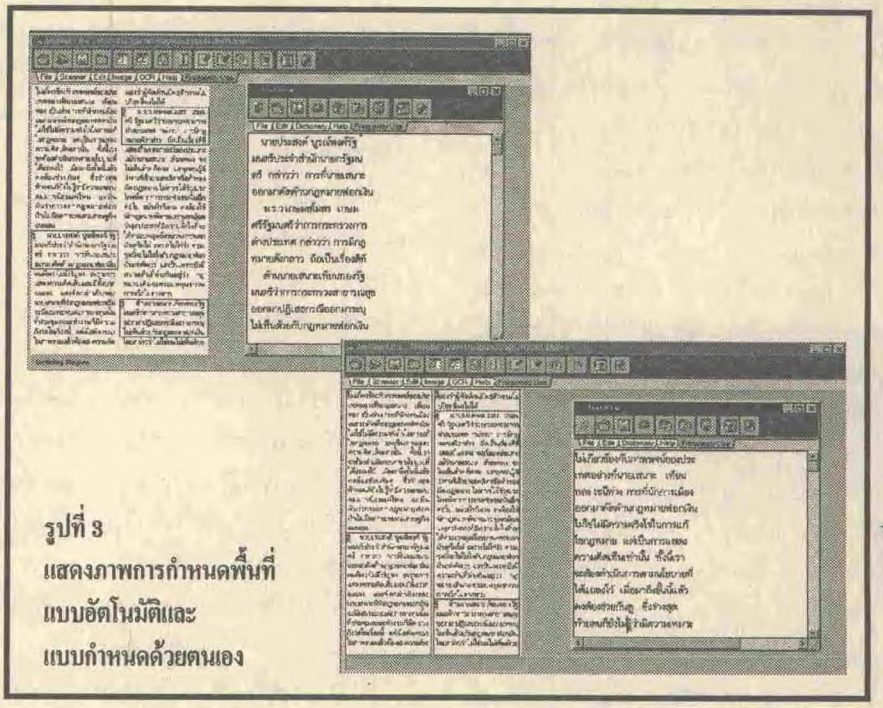
3.ขบวนการหลังประมวลผล (Postprocessing)

ขั้นที่ 1 ขบวนการเตรียมประมวลผลนั้นเป็นการรับภาพลักษณ์เอกสารเข้ามา ซึ่งอาจรับมาโดยการติดต่อเข้ากับเครื่องกราดตรวจโดยตรงก็ได้ หรือนำแฟ้มภาพลักษณ์มาเปิดก็ได้ ซึ่งแฟ้มภาพลักษณ์ในตระกูลต่อไปนี้สามารถถูกเปิดได้ คือ Ventura IMG (.IMG), Windows Metafile (.WMF), Word Perfect Graphic (.WPG), encapsulated PostScript (.EPS), Windows BMP และ DIB (.BMP, .DIB), TIFF (.TIF), JPEG (.JPG), PCX (.PCX), และ Targa (.TGA) หลังจากนั้นผู้ใช้พิจารณาดูว่าภาพลักษณ์ที่นำเข้ามานั้นมีความเอียงจากการสแกนหรือไม่ ถ้ามีก็ควรจะสั่งให้ซอฟต์แวร์ปรับภาพลักษณ์ให้ตั้งตรง ดังรูปที่ 2 ซึ่งการทำให้ภาพลักษณ์ตั้งตรงนี้ผู้ใช้ไม่จำเป็นต้องทราบองศาของความเอียงของเอกสาร ซอฟต์แวร์จะทำการตรวจสอบเอง นอกจากนี้ถ้าผู้ใช้ต้องการปรับแต่งภาพ โดยการกลับซ้ายขวา กลับหัวกลับท้าย หรือกลับด้าน ก็สามารถทำได้เช่นกัน

นอกจากนี้ผู้ใช้ยังสามารถย่อ ขยาย หรือปรับความคมชัดของภาพลักษณ์เพื่อให้สามารถอ่านข้อความในภาพลักษณ์ได้ชัดเจนยิ่งขึ้นได้ โดยจะอนุญาตให้มีระยะของการ Zoom-In เป็น 125%, 150%, 175%, และ 200% และระยะของการ Zoom-Out เป็น 75%, 50%, และ 25% และเมื่อท่านปรับขนาดภาพที่แสดงเป็นที่พอใจแล้ว ขั้นต่อไปจะเป็นการกำหนดพื้นที่ในภาพลักษณ์เพื่อนำไปทำการรู้จำตัวอักษรซึ่งมีทั้งแบบกำหนดพื้นที่โดยอัตโนมัติเหมาะกับเอกสารที่มีหลาย ๆ คอลัมน์และหลายย่อหน้า และแบบกำหนดด้วยตัวเอง ซึ่งจะกำหนดพื้นที่ขนาดเท่าใดและจำนวนเท่าไรก็ได้ ดังแสดงในรูปที่ 3 แต่ถ้าท่านต้องการทำการรู้จำทั้งภาพ ท่านก็ข้ามขั้นตอนนี้ไปได้เลย



รูปที่ 2
แสดงการปรับตรงของภาพลักษณ์



รูปที่ 3
แสดงภาพการกำหนดพื้นที่แบบอัตโนมัติและแบบกำหนดด้วยตนเอง

ขั้นที่ 2 เป็นหัวใจของการทำงาน คือการอ่านตัวอักษรหรือการรู้จำตัวอักษรไทยและอังกฤษนั่นเอง ซึ่งในซอฟต์แวร์อ่านไทยรุ่นใหม่ที่จะออกสู่ท้องตลาดนี้จะเรียนรู้จำตัวอักษรไทยและอังกฤษของฟอนต์ไทยจำนวน 20 ฟอนต์ ซึ่งเป็นฟอนต์ที่นิยมใช้ในเครื่องคอมพิวเตอร์ส่วนบุคคล (PC) และฟอนต์ที่นิยมในเครื่องแมคอินทอช ซึ่งแตกต่างจากซอฟต์แวร์อ่านไทยรุ่น 1.0 ซึ่งมีให้เพียง 10 ฟอนต์และเป็นฟอนต์ที่นิยมใช้บนเครื่องคอมพิวเตอร์ส่วนบุคคลทั้งสิ้น และในรุ่น 1.0 ถ้าผู้ใช้ทราบว่าเอกสารที่นำมาสแกนใช้นั้นเป็นฟอนต์ใดจะเป็นการดีที่สุด เพราะจะทำให้ผลลัพธ์ที่ได้มีความถูกต้องมากที่สุด ซึ่งไม่อำนวยความสะดวกในการใช้งานให้กับผู้ใช้มากนัก เพราะผู้ใช้มักไม่ทราบว่าเอกสารที่มีอยู่นั้นพิมพ์ด้วยฟอนต์ใด ถึงแม้จะได้มีการเตรียมส่วนสำหรับให้ผู้ใช้พิมพ์คำที่มีในภาพลักษณ์เพื่อใช้ในการเปรียบเทียบฟอนต์ให้แล้ว ก็ยังเป็นการยากที่จะใช้งานอยู่ดี ดังนั้นในรุ่นใหม่นี้ ผู้ใช้จะไม่ต้องระบุฟอนต์อีกต่อไปแล้ว ขอเพียงท่านระบุว่าเอกสารที่ได้มานี้พิมพ์จากเครื่องคอมพิวเตอร์ส่วนบุคคลหรือ

เรื่องน่าอ่าน

ซอฟต์แวร์ไทยโอซีอาร์
: พนักงานพิมพ์ดีดคนใหม่

เครื่องแมคอินทอชก็เพียงพอแล้ว ซึ่งจะอำนวยความสะดวกให้แก่ผู้ใช้งานมากยิ่งขึ้น อัตรารู้จำโดยเฉลี่ยของซอฟต์แวร์อ่านไทย รุ่นใหม่นี้สำหรับเอกสารผสมทั้งภาษาไทยและภาษาอังกฤษนั้นจะอยู่ที่ 95% หรือสูงกว่า ซึ่งเมื่อเทียบกับรุ่น 1.0 จะอยู่ที่เพียง 90% เท่านั้น

เทคโนโลยีของการรู้จำที่ใช้ในซอฟต์แวร์อ่านไทยนี้ พัฒนาโดยคณะผู้วิจัยจากหน่วยปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ประกอบด้วย ดร. ทวีศักดิ์ กออนันตกูล (ผู้เชี่ยวชาญเรื่องภาษาไทยกับคอมพิวเตอร์และเรื่องอินเทอร์เน็ต) เป็นที่ปรึกษาโครงการ ดร. จุฬารัตน์ ตันประเสริฐ เป็นหัวหน้าโครงการ และนักวิจัยประกอบด้วย คุณวศิน สินธุภิญโญ คุณเปรมนาถ ดูเบ และคุณสุทัศน์ แซ่ตั้ง (น้องใหม่ของโครงการ) เทคโนโลยีที่คณะผู้วิจัยนำมาใช้ในการรู้จำตัวอักษรคือโครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นสาขาที่พยายามทำให้คอมพิวเตอร์ทำงานบางอย่างที่ยังทำไม่ได้ ให้ได้เหมือนมนุษย์ อย่างเช่น งานการจัดแยกสิ่งของ งานรู้จำตัวอักษร หรืองานรู้จำเสียงพูด เป็นต้น คุณลักษณะเด่นของเทคโนโลยีนี้ประกอบด้วย

- มีความยืดหยุ่นสูง สามารถจำลองขบวนการของปัญหาใด ๆ ได้
- มีความสามารถในการจำคู่อินพุต-เอาต์พุตที่มีความซับซ้อนมาก ๆ ได้
- มีความสามารถในการปรับตัวเข้ากับการเปลี่ยนแปลงของสิ่งแวดล้อมได้
- มีความสามารถในการตอบสนองต่อข้อมูลที่ไม่เคยเห็นมาก่อนได้
- ความรู้กระจายอยู่ทั่วทั้งโครงสร้างของโครงข่ายประสาทเทียม ส่งผลให้มีความทนทานต่อความผิดพลาดหรือความล้มเหลวขององค์ประกอบของโครงข่ายได้มาก

ซึ่งคุณสมบัติเด่นเหล่านี้มีความเหมาะสมกับลักษณะงานของไทยโอซีอาร์เป็นอย่างยิ่ง อาทิเช่น ความสามารถในการตอบสนองต่อข้อมูลที่ไม่เคยเห็นมาก่อน ในการสแกนเอกสารหน้าเดิม แต่สแกนคนละครั้ง จะพบว่า

ภาพที่ได้ไม่เหมือนกัน มีบางจุด (pixel) ที่ได้จะสลับขาวเป็นดำ หรือดำเป็นขาว แต่ภาพโดยรวมของภาพลักษณ์แต่ละตัวอักษรจะยังคงอ่านได้ด้วยตามนุษย์ และเช่นกัน โครงข่ายประสาทเทียมก็มีความสามารถตรงจุดนี้ สามารถอ่านภาพลักษณ์ที่สแกนเข้ามาใหม่และไม่เหมือนกับภาพลักษณ์ที่เคยสอนไว้แต่มีลักษณะคล้ายกันได้ ในปัจจุบันนี้มีแบบจำลอง (Model) ของโครงข่ายประสาทเทียมอยู่มากมาย ซึ่งแต่ละแบบจำลองจะเหมาะสมกับการใช้งานต่าง ๆ กันไป และสร้างขึ้นมาจากแนวความคิดความเข้าใจที่ต่างกัน เพื่อที่จะเลียนแบบการทำงานของสมองของมนุษย์ แต่เนื่องจากยังไม่มีนักวิทยาศาสตร์คนใดสามารถสร้างแบบจำลองการทำงานจริง ๆ ของสมองมนุษย์ออกมาได้ ทางนักวิจัยฝ่ายระบบคอมพิวเตอร์ก็ได้แต่คาดการณ์เอาว่าน่าจะเป็นแบบนั้นแบบนี้ จึงปรากฏออกมาเป็นแบบจำลองมากมาย อาทิเช่น แบบจำลองแพร่กระจายย้อนหลัง (Back-Propagation Model) แบบจำลองการจัดแผนที่ลักษณะเด่นแบบโคโฮเน็น (Kohonen's Self-Organizing Feature Map Model) แบบจำลองนีโอค็อก-นีตรอน (Neocognitron) และแบบจำลองทฤษฎีการกำทอนแบบปรับตัวได้ (Adaptive Resonance Theory : ART) เป็นต้น ถ้าท่านผู้อ่านสนใจแนวคิดและแบบจำลองของโครงข่ายประสาทเทียม ผู้เขียนขอแนะนำหนังสือน่าอ่าน 2 เล่มได้แก่ "Fundamentals of Neural Networks : Architectures, Algorithms, and Applications" โดย Laurene Fausett และหนังสือ "Neural Networks in Computer Intelligence" โดย Limin Fu เพื่อเป็นการปูพื้นฐานความเข้าใจ

สำหรับแบบจำลองที่ทางคณะผู้วิจัยนำมาใช้ในการพัฒนาโอซีอาร์ประกอบด้วย 2 แบบจำลอง คือ แบบจำลองแพร่กระจายย้อนหลัง (Back-Propagation Model) และแบบจำลองการจัดแผนที่ลักษณะเด่นแบบโคโฮเน็น (Kohonen's Self-Organizing Feature Map Model) ซึ่งเป็นแบบจำลองที่ได้รับความนิยมมากในการนำไปประยุกต์ใช้งานต่าง ๆ ทางคณะผู้วิจัยได้นำแบบจำลองโคโฮเน็น (ขอเรียกสั้น ๆ เพื่อลดความสับสน) มาใช้ในการแบ่งกลุ่มตัวอักษรให้ตัวอักษรที่มีลักษณะคล้ายคลึงกันในแง่มุมต่าง ๆ มาอยู่ร่วมกัน แบบจำลองนี้เป็นแบบไม่มีครูสอน นั่นคือเราไม่ทราบว่าควรจะจัดตัวอักษรใดอยู่ร่วมกัน หรือควรจะนำเกณฑ์ใดมาใช้ในการแบ่งกลุ่มตัวอักษร แบบจำลองนี้จะทำการหาหลักเกณฑ์เองและแบ่งกลุ่มตัวอักษรออกมาให้ ผลที่ได้จะส่งให้แบบจำลองแพร่กระจายย้อนหลังไปฝึกหัดให้แต่ละโครงข่ายรู้จำตัวอักษรในแต่ละกลุ่มอีกทอดหนึ่ง ซึ่งแบบจำลองนี้เป็นแบบมีครูสอน นั่นคือทุก ๆ อินพุต เราจะใส่เอาต์พุตไปด้วยว่าคืออะไร เพื่อจะให้เป็นครูสอนแก่โครงข่ายประสาทเทียมว่าหน้าตาอินพุตนี้คือเอาต์พุตนี้นะ โครงข่ายจะทำการเรียนรู้รอบแล้วรอบเล่าจนกว่าจะจำได้ เปรียบไปก็คล้ายกับการสอนให้เด็กอ่านหนังสือ เริ่มด้วยสอน "ก" "ข" และ ตัวอักษรอื่น ๆ เรื่อย ๆ ไป แรก ๆ เด็กก็จำได้มั่งไม่ได้มั่ง แต่พอนานวันเข้าเด็กก็จำได้แม่น และเมื่อเห็นตัวอักษรที่เขียนแบบประดิษฐ์ หรือแบบไม่เรียบร้อย เด็กก็ยังอ่านได้ว่าคืออักษรใด นั่นคือมีความสามารถในการเรียนรู้ได้ ซึ่งเป็นความสามารถหนึ่งของโครงข่ายประสาทเทียมเช่นกัน

ซอฟต์แวร์ไทยโอซีอาร์
: พนักงานพิมพ์ดีดคนใหม่

การใช้งานโครงข่ายประสาทเทียมถูกแบ่งออกเป็น 2 ส่วนหลัก คือส่วนการฝึกหัดโครงข่าย และส่วนการนำไปใช้งาน ส่วนที่ผู้ใช้ซอฟต์แวร์อ่านไทยได้สัมผัสคือส่วนที่สอง สำหรับส่วนแรกนั้นทางคณะผู้วิจัยได้ทำการสอนหรือฝึกหัดให้โครงข่ายได้รู้จำตัวอักษรไว้ก่อนแล้ว และค่อยนำผลลัพธ์ที่ได้ไปใช้งาน ในการฝึกหัดนั้นเราสร้างข้อมูลจำนวนมาก เพื่อให้โครงข่ายรู้จักรูปแบบตัวอักษรหลากหลายรูปแบบ เสมือนการสอนให้เด็กได้เห็น "ก" หลาย ๆ แบบนั่นเอง จำนวนตัวอักษรทั้งหมดที่ส่งไปฝึกหัดนั้นประกอบด้วย ตัวอักษรที่แตกต่างกันทั้งสิ้น 180 ตัวอักษร ในแบบตัวธรรมดา ตัวหนา ตัวเอียง และตัวหนาเอียง และในขนาดตัวอักษรแตกต่างกัน 6 ขนาด ของฟอนต์ที่แตกต่างกันทั้งหมด 20 ฟอนต์ และการสแกนด้วยระดับความละเอียดแตกต่างกัน 3 ระดับ ดังนั้นข้อมูลทั้งหมดที่ใช้ในการฝึกหัดโครงข่ายประสาทเทียมคือ 180x4x6x20x3 = 307,200 ตัวอักษร ซึ่งเป็นปริมาณข้อมูลที่มากทีเดียว

ในการฝึกหัดนี้เราใช้ซอฟต์แวร์ที่มีชื่อว่า SNNS ชื่อเต็มว่า Stuttgart Neural Network Simulator ซึ่งเป็นซอฟต์แวร์สำหรับการฝึกหัดโครงข่ายประสาทเทียมหลาย ๆ แบบจำลอง พัฒนาโดยมหาวิทยาลัย Stuttgart เผยแพร่ให้ใช้เพื่องานวิจัยที่เกี่ยวเนื่องกับโครงข่ายประสาทเทียม ใครที่สนใจอยากจะลองเอามาใช้ดู เข้าไปค้นข้อมูลเพิ่มเติมได้ที่

```
URL http://ftp.informatik.uni-
        stuttgart.de/pub/SNNS/
```

ซอฟต์แวร์นี้ทำงานบนระบบปฏิบัติการ UNIX หรือ Linux ถ้าท่านไม่ได้ใช้ระบบปฏิบัติการดังกล่าว ลองเข้าไปดูซอฟต์แวร์อื่น ๆ ที่มีให้ดาวน์โหลดใช้ได้ที่

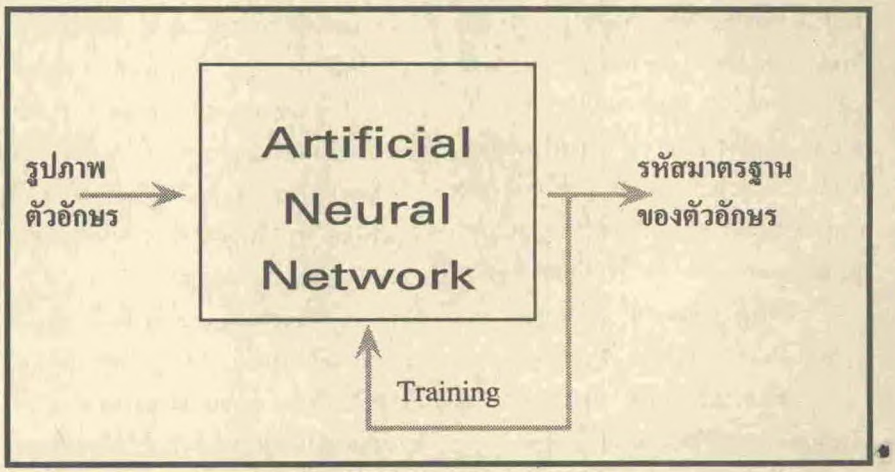
```
URL http://www.neuronet.ph.kcl.ac.uk/
        neuronet/software/software.html
```

มีข้อมูลทั้งซอฟต์แวร์ที่พัฒนาเพื่อจำหน่ายและเพื่อแจกจ่ายด้านเทคโนโลยีโครงข่ายประสาทเทียมอยู่มากมาย และนอกจากนี้เรายังได้อาศัยศักยภาพของเครื่อง SGI Power Challenge XL ซึ่งบรรจุ 8 R10000 หน่วยประมวลผล (processors), 1 GB หน่วยความจำ (RAM) และ 50 GB ฮาร์ดดิสก์ ของหน่วยปฏิบัติการวิจัยและพัฒนาเทคโนโลยีคอมพิวเตอร์สมรรถนะสูง สังกัดเนคเทคอีกด้วย เพราะการฝึกหัดข้อมูลขนาดใหญ่มากนี้ หน่วยความจำในเครื่องคอมพิวเตอร์ส่วนบุคคลที่ทางคณะผู้วิจัยมีอยู่สูงสุด 64 MB ไม่สามารถบรรจุได้หมด การฝึกหัดใช้เวลาเป็นวัน ๆ มากน้อยแล้วแต่ขนาดของอินพุตและจำนวนตัวอักษรที่ส่งเข้าไปฝึกหัด มีปัจจัยที่ทางคณะผู้วิจัยต้องทำการศึกษาหลายปัจจัย ซึ่งในปัจจุบันนี้ได้คำตอบทุกด้านแล้ว ปัจจัยเหล่านั้นได้แก่

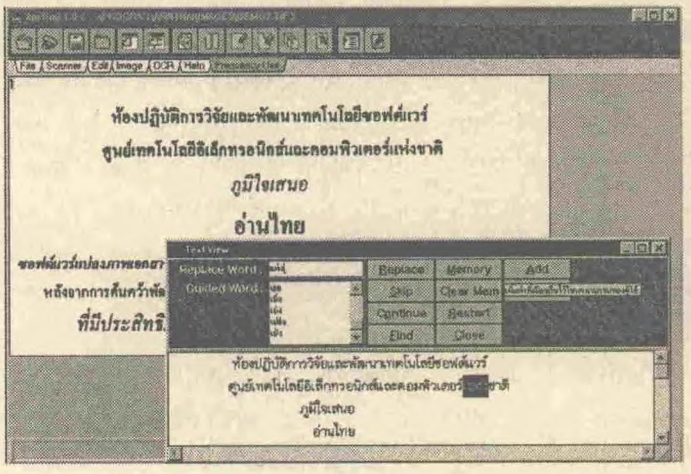
- รูปแบบและขนาดของอินพุตที่จะนำเข้ามาทำการเรียนรู้ใน ANN
- แบบจำลองและเทคนิคการเรียนรู้ของ ANN
- การปรับแต่งสถาปัตยกรรมโครงข่ายเพื่อให้มีขีดความสามารถในการเรียนรู้จำตัวอักษรไทยได้ดีที่สุด
- จำนวนตัวอย่างข้อมูลที่จะนำมาฝึกหัดให้โครงข่ายรู้จัก

ขั้นที่ 3 เป็นส่วนการประมวลผลหลังการรู้จำ ซึ่งประกอบไปด้วย 2 ขบวนการหลัก คือ

1. การใช้กฎของการรวมคำไทยมาใช้เพื่อช่วยในการระบุว่าเอาต์พุตภาพลักษณ์นี้ควรจะเป็นตัวอักษรใด
2. การใช้พจนานุกรมมาช่วยตรวจสอบคำสะกดอีกต่อหนึ่ง



รูปที่ 4



รูปที่ 5 แสดงผลการแปลงภาพลักษณ์ของเอกสารเป็นแฟ้มข้อความและการตรวจสอบคำสะกด

สำหรับขบวนการแรกนั้นทางคณะผู้วิจัยได้ทำการศึกษาถึงกฎการรวมคำตามหลักภาษาไทย แล้วพัฒนาใส่เป็นกฎไว้ในโปรแกรม ตัวอย่างเช่น "ฤา" ต้องแก้เป็น "ฤๅ" หรือ "ปฺสน" แก้ไขเป็น "เป็น" เป็นต้น ส่วนการตรวจสอบคำสะกดนี้จะตรวจสอบได้ทั้งภาษาไทยและภาษาอังกฤษในขั้นตอนเดียว เมื่อซอฟต์แวร์ตรวจพบคำที่สะกดผิด ผู้ใช้จะสามารถเลือกที่จะแก้ไขคำให้ถูกต้องได้ 2 วิธีคือ เลือกคำจากรายการคำที่แสดงให้ หรือพิมพ์คำที่ต้องการใส่ลงไปก็ได้ นอกจากนี้ผู้ใช้ยังสามารถทำให้ซอฟต์แวร์อ่านไทยรู้จักคำศัพท์มากขึ้นได้ โดยการเลือกคำที่ต้องการและกดปุ่ม "Add" เพื่อนำไปจัดเก็บไว้ในพจนานุกรมของผู้ใช้ได้อีกด้วย

สำหรับท่านผู้อ่านที่สนใจอยากได้รายละเอียดเพิ่มเติมเกี่ยวกับซอฟต์แวร์อ่านไทย สามารถติดต่อได้ที่บริษัท ไทยซอฟท์ จำกัด หรือถ้าเป็นข้อมูลทางเทคนิคเกี่ยวกับ Thai & English OCR Engine ก็สามารถดูรายละเอียดเพิ่มเติมได้ที่

```
http://swl.nectec.or.th/thaiocr/
```

หรือ โทร. 642-7059..63 ต่อ 6001-6006 และสำหรับท่านผู้อ่านที่สนใจอยากจะทดสอบซอฟต์แวร์นี้ก็สามารถทำได้ โดยไปที่โฮมเพจข้างต้นและไปที่หัวข้อ Software Testing และทำตามขั้นตอนที่แนะนำไว้ ท่านจะสามารถ Upload แฟ้มภาพลักษณ์ในรูปแบบ ".BMP" ที่สแกนแบบขาวดำ มาที่ Server และทาง Server จะแปลงแฟ้มนั้นให้เป็นแฟ้มข้อความด้วย Thai & English OCR Engine และส่งแฟ้มข้อความออกเป็นเอาต์พุตที่หน้าจอของเครื่องคอมพิวเตอร์ของท่าน

อย่างที่ท่านผู้อ่านทราบกันว่าฟอนต์ภาษาไทย (ซึ่งรวมอักษรภาษาอังกฤษไว้ด้วย) นั้นมีอยู่มากมาย หน้าตาแตกต่างกันออกไปมากบ้างน้อยบ้าง บางฟอนต์ตัวอักษรเรียบง่าย บางฟอนต์เป็นแบบประดิษฐ์เพื่อความสวยงาม ซึ่งส่งผลให้ซอฟต์แวร์ไทยโอซีอาร์จะต้องมีความสามารถสูงมากในการที่จะเข้าใจและอ่านภาพลักษณ์ตัวอักษรไทยได้ทุกตัวและทุกฟอนต์ สำหรับซอฟต์แวร์อ่านไทยเลือกที่จะทำงานเฉพาะกับฟอนต์ภาษาไทยที่ใช้ในการเอกสาร จดหมาย หรือหนังสือทั่ว ๆ ไป โดยไม่ทำการรู้จำที่จะอ่านอักษรในฟอนต์ประดิษฐ์เลย ถ้าประเทศไทยมีฟอนต์มาตรฐานที่ว่า ไม่ว่าเครื่องระบบใด ๆ ก็จะมีหน้าตาเหมือนกัน หรือมีมาตรฐานในการออกแบบฟอนต์ ก็จะเป็นประโยชน์ต่อการพัฒนาซอฟต์แวร์ไทยโอซีอาร์เป็นอย่างยิ่ง เท่าที่ทราบขณะนี้ทางศูนย์เทคโนโลยีอิเล็กทรอนิกส์ได้เป็นผู้ประสานงานระหว่างผู้เชี่ยวชาญจากหน่วยงานหลาย ๆ หน่วยงาน เพื่อเป็นคณะกรรมการหาแนวทางการกำหนดมาตรฐานต่าง ๆ เกี่ยวกับฟอนต์ภาษาไทยกันอยู่ ก็ขอเป็นกำลังใจให้ทำงานสำเร็จบรรลุเป้าหมายได้ด้วยดี เพื่อประโยชน์ต่อประเทศไทยของเรา

เอกสารอ้างอิง

1. คู่มือการใช้งานโปรแกรมอ่านไทย เวอร์ชัน 1.0 ลิขสิทธิ์ พ.ศ. 2539

2. Proceedings of the IEEE, Specia, Issue on Optica, Character Recognition, July 1992, Vo. 80, No. 7

3. Proceedings of the Symposium on Natura, Language Processing in Thai, and, March 1993.

คุณลักษณะของซอฟต์แวร์อ่านไทย รุ่น 1.5

- ทำการแปลงรูปภาพเอกสารตัวพิมพ์ไทยและอังกฤษให้เป็นแฟ้มข้อความ
- รับอินพุตที่เป็นแฟ้มภาพลักษณ์เอกสารแบบ Ventura IMG (.IMG), Windows Metafile (.WMF), Word Perfect Graphic (.WPG), encapsulated PostScript (.EPS), Windows BMP และ DIB (.BMP, .DIB), TIFF (.TIF), JPEG (.JPG), PCX (.PCX), และ Targa (.TGA)
- ส่งเอาต์พุตเป็นแฟ้มข้อความเอกสารแบบ ".txt"
- สามารถทำงานได้กับเอกสารที่สแกนเข้ามาเอียง เนื่องจากความผิดพลาดของผู้สแกนหรือเครื่องกราดตรวจ ได้อย่างอัตโนมัติ
- มีส่วนการย่อ/ขยายภาพลักษณ์เอกสารที่มีประสิทธิภาพ รวมถึงส่วนแสดงผลลัพธ์ที่ได้ในแต่ละขั้นตอนการทำงานอย่างชัดเจน
- อัตราความถูกต้องในการรู้จำเฉลี่ยที่ 95% กับเอกสารภาษาไทยและภาษาอังกฤษตัวพิมพ์ในฟอนต์ไทยจำนวน 20 ฟอนต์ ซึ่งเป็นฟอนต์ที่นิยมใช้ในเครื่อง PC และเครื่อง Macintosh
- ใช้งานได้กับต้นฉบับที่มีขนาดตัวอักษรตั้งแต่ 12 points ขึ้นไป
- มีส่วนการเลือกทำงานเฉพาะบางส่วนของหน้ารูปภาพเอกสาร หรือกำหนดพื้นที่โดยอัตโนมัติได้
- ทำงานได้ดีกับภาพลักษณ์แบบขาวดำ ที่สแกนเข้ามาด้วยระดับความละเอียด 200-400 dpi
- มีส่วนตรวจสอบคำสะกด เพื่อเพิ่มความถูกต้องในกับแฟ้มข้อความอีกด้วย